	*QMRF identifier (JRC Inventory):***To be entered by JRC**
	*QMRF Title:***MLog P model v. 1.0.0 in VEGA v. 1.1.4**
	*Printing Date:***19-feb-2020**

1.QSAR identifier

1.1.QSAR identifier (title):

MLog P model v. 1.0.0 in VEGA v. 1.1.4

1.2.Other related models:

1.3.Software coding the model:

VEGA v. 1.4.4

https://www.vegahub.eu/portfolio-item/vega-qsar/

2.General information

2.1.Date of QMRF:

11 April 2010

2.2.QMRF author(s) and contact details:

[1]Domenico Gadaleta Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri", IRCCS domenico.gadaleta@marionegri.it

[2]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri", IRCCS emilio.benfenati@marionegri.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Ikuo Moriguchi School of Pharmaceutical Sciences, Kitasato University, 5-9-1 Shirokane, Minato-ku, Tokyo 108, Japan.

2.6.Date of model development and/or publication:

1992

2.7.Reference(s) to main scientific papers and/or software package:

[1]I.Moriguchi, S.Hirono, Q.Liu, I.Nakagome, and Y.Matsushita, Chem.Pharm.Bull. 1992, 40, 127-130

[2]I.Moriguchi, S.Hirono, I.Nakagome, H.Hirano, Chem.Pharm.Bull. 1994, 42, 976-978.

2.8.Availability of information about the model:

Model’s guide is available for download from VEGA v. 1.1.4

2.9.Availability of another QMRF for exactly the same model:

Other QMRF for this model are not available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

N/A

3.2.Endpoint:

QMRF 1. 6. Octanol-water partition coefficient (Kow) EC A.8 Partition Coefficient (EU method includes both shake flask and HPLC)

3.3.Comment on endpoint:

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

Logarithm of octanol/water partition coefficient (log P)

3.6.Experimental protocol:

EC A.8 Partition Coefficient

OECD 123 Partition Coefficient (nOctanol/Water): Slow-Stirring Method

OECD 117 Partition Coefficient (n-octanol/water) HPLC Method

OECD 107 Partition Coefficient (noctanol/water); Shake Flask Method

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Regression equation based on structural parameters

4.2.Explicit algorithm:

Linear regression

The MlogP models in VEGA 1.4.4 implement the multiple linear regression developed by Morigouchi et al. (1992; 1995) that relates 13 sructural parameters with the experimental log P values of 1230 compounds with different structures and including C, H, N, O, S, P, F, Cl, Br. The equation is the following:

MlogP = -1.041 + 1.244(CX) $^{0.6}$- 1.017(NO) $^{0.9}$+
0.406(PRX) - 0.145(UB) $_{0.8}$+ 0.511(HB) + 0.268(POL) -
2.215(AMP) + 0.912(ALK) -0.392(RNG) -3.684(QN) + 0.474(NO2) + 1.582(NCS)
+ 0.773(BLM)

The model variables are frequencies (denoted by N) or presence/absence (denoted by D) of some molecular features. Their description is reported in the table below:

Parameter (Description)

C (Summation of weighted numbers of carbon and halogen atoms; the weights are: 0.5 for F, 1.0 for C and Cl, 1.5 for Br, and 2.0 for I).

NO (Total number of Ns and Os).

PR (Proximity effect of N/O: 2 for X-Y and 1 for X-A-Y (X, Y: N and/or O; A: C, S, or P; -: saturated or unsaturated bond) with a correction (-1) for -CON < and -SO2N <)

U (Number of unsaturated bonds including semi-polar bonds such as N-oxides and sulfoxides, except those in NO2=).

H (Dummy variable for the presence of intramolecular hydrogen bond as ortho-OH and -CO-R, -OH and -NH2, -NH2 and -COOH, or 8-OH/NH2 in quinolines, 5 or 8-OH/NH2 in quinoxalines, etc.)

POL (Number of aromatic polar substituents (aromatic substituents excluding Ar-C(X)(Y)- and Ar-C(X)=C; X, Y: C and/or H). Upper limit = 4)

AMP (Amphoteric property; a-aminoacid = 1, aminobenzoic acid = 0.5, pyridinecarboxylic acid = 0.5).

ALK (Dummy variable for alkane, alkene, cycloalkane, cycloalkene (hydrocarbons with 0 or 1 double bond) or hydrocarbon chain with at least 7 carbon atoms).

RNG (Dummy variable for the presence of ring structures except benzene and its condensed rings (aromatic, heteroaromatic, and hydrocarbon rings)

QN (Quaternary nitrogen >N+ <: 1; N-oxide: 0.5).

NO2 (Number of nitro groups).

NCS (Isothiocyanate (-N=C=S): 1.0; thiocyanate (-S-C#N): 0.5).

BLM (Dummy variable for the presence of ß-lactam)

4.3.Descriptors in the model:

The model variables are frequencies (denoted by N) or presence/absence (denoted by D) of some molecular features. Their description is reported in the tableabove (see 4.2)

4.4.Descriptor selection:

Details on the selection procedure of the 13 parameters and on the development of the multiple linear regression are reported in Morigouchi et al., 1992.

4.5.Algorithm and descriptor generation:

See 4.2

4.6.Software name and version for descriptor generation:

4.7.Chemicals/Descriptors ratio:

1200 / 13 = 92.3

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model implemented in VEGA v. 1.4.4 is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments). For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation. Following, all applicability domain components are reported along with their explanation and the intervals used.

· Similar molecules with known experimental value. This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

- 1 >= index > 0.9 strongly similar compounds with known experimental value in the training set have been found.
- 0.9 >= index > 0.75 only moderately similar compounds with known experimental value in the training set have been found.
- index <= 0.75 no similar compounds with known experimental value in the training set have been found.

· Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are:- index < 0.5 accuracy of prediction for similar molecules found in the training set is good - 0.5 <=

index < 1.0 accuracy of prediction for similar molecules found in the training set is not optimal.- index > 1.0 accuracy of prediction for similar molecules found in the training set is not adequate.

· Concordance with similar molecules (average difference between target compound prediction and experimental values of similar molecules) . This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made agrees with the experimental values found in the model's space, thus the prediction is reliable. Defined intervals are:- index < 0.5 similar molecules found in the training set have experimental values that agree with the target compound predicted value.- 0.5 <= index < 1.0 similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value.- index > 1.0 similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value.

- · Maximum error of prediction among similar molecules. This index takes into account the maximum error in prediction among the two most similar compounds. Values near 0 means that the predicted compounds falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

- index < 0.5 the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability.- 0.5 <= index < 1.0 the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability.

- index >= 1.0 the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability.

· Global AD Index. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are:

- 1 >= index > 0.85 predicted substance is into the Applicability Domain of the model.

- 0.85 >= index > 0.75 predicted substance could be out of the Applicability Domain of the model.- index <= 0.75 predicted substance is out of the the Applicability Domain of the model.

5.2.Method used to assess the applicability domain:

5.3.Software name and version for applicability domain assessment:

VEGA v. 1.4.4

https://www.vegahub.eu/portfolio-item/vega-qsar/

5.4.Limits of applicability:

The model was developed starting from a dataset of structures including only C, H, N, O, S, P, F, Cl, Br elements.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes

Chemical Name: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3.Data for each descriptor variable for the training set:

No

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

The 1200 compounds used by Morigouchi et al. to derive the multiple linear regression were cited from Hansh & Leo, “Substituent Constant for Correlation Anaysis in Chemistry and Biology”, John Wiley and Sons, New York, 1979.

The training set of the Meylan LogP model (9,961 compounds) was used as training set during the implementation.

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

Performance on the training set are reported in Morigouchi et al., 1992:

n = 1230, r = 0.952, s = 0.411, $F_0(13,1216) = 900.4$

On the pruned training set from EPI Suite KowWin module (9,961 compounds), the logP model has the following statistics:

Test set: n = 9961; R2 = 0.73; RMSE = 0.96

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

7.3.Data for each descriptor variable for the external validation set:

Unknown

7.4.Data for the dependent variable for the external validation set:

Unknown

7.5.Other information about the external validation set:

7.6.Experimental design of test set:

7.7.Predictivity - Statistics obtained by external validation:

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The two main parameters in the equation are CX (i.e., the summation of empirically weighted numbers of carbon and halogen atoms) accounting the hydrophobic contribution to log P, and NO (i.e. the total number of nitrogen and oxygen atoms) accounting for the hydrophilic contribution to log P.

The proximity effect of nitrogen and oxygen atoms was also considered important as a correction for the electronic structures and implemented with PRX parameter.

The remaining parameters account for the effect of various substructures.

8.2.A priori or a posteriori mechanistic interpretation:

A priori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]I.Moriguchi, S.Hirono, Q.Liu, I.Nakagome, and Y.Matsushita, Chem.Pharm.Bull. 1992, 40, 127-130

[2]I.Moriguchi, S.Hirono, I.Nakagome, H.Hirano, Chem.Pharm.Bull. 1994, 42, 976-978.

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC